Name: _____ Date: _____

A2CC: Intro to Statistics

Data is everywhere. It's in our newspapers, it's in our science classes, it shows up in economics, medicine and anywhere else that **variability** occurs. Variability is simply the property of **outcomes** being different. The tools of statistics are designed to **explain this variability**.

There are many types of variability. It is good to understand these sources in order to minimize the ones that we are not studying.

(a) **Observational or Measurement Variability**: Variability that is introduced due to either our measuring instruments not being precise enough or differences in how two different people read the measurement.

(b) **Natural Variability or Inter-Individual Variability**: Variability that accounts for the fact that members of a populations are simply different.

(c) **Induced Variability**: This type of variability is in marked contrast to natural. It occurs because we have assigned our population or sample to two or more **treatment** groups and then observe the variability between the groups.

(d) **Sample Variability**: This is the type of variability that occurs when we take multiple **samples** from a **population** randomly. These samples will be different due to the randomness of the sampling process.

Remember, through all of our work in this unit, we are really trying to explain the variability of data within either a population or a sample and then using this to determine if the variability can be attributed to one of the factors above to the exclusion of the others.

There are many different situations in which we collect data. They have important differences and all of them depend on **randomization** in one way or another.

The three major types of ways to collect data are described below. Let's give an example of each and explain how **randomization** is part of each method. Randomization is used primarily to eliminate variability caused by some type of **bias**.

(a) **Surveys**: Collections of data from a population where variability is not induced by treatments but by the sample itself (sampling variability).

(b) **Observational Studies**: Collections of data from a population where assignment of individuals from the population into **treatment groups** is **not** under the control of those performing the study.

(c) **Experimental Studies**: In experimental studies individuals are assigned randomly to treatment groups in order to determine the effect of the treatment on the variability of the data. In these cases, the assignment, although random, is under the control of those performing the study.

Random sampling is critical for being able to minimize variability due to **sampling bias**. Random sampling can be done using a variety of different techniques. Simple random sampling can be accomplished using a random number table.

***Exercise* #1:** A list of 10 people's heights, in inches, is shown below.

Person #	1	2	3	4	5	6	7	8	9	10
Height	70	68	60	75	65	69	58	62	66	63

(a) Randomly select five heights from this list by using the random number table that goes with this lesson. Choose a random spot in the table and move down the column. Select the first digit of each number. If you get a repeat, eliminate and keep going. If you get a 0, use this as the 10.

(b) Calculate the **sample mean** to the nearest tenth. Compare to others in the class. What type of variability is being introduced through this process?

When we conduct a study, the complete set of all subjects that share a common characteristic that is being studied is known as the **population**. All populations have **natural or inter-individual variability**. Most of the time, the entire population is not measured, but a sample is taken to infer characteristics of a population. Still, all populations in theory have **population parameters** that describe the population, such as its mean, standard deviation, and interquartile range.

***Exercise* #2:** 18 students in Mr. Weiler's Advanced Calculus class took a quiz with the following results in ascending order.

56, 68, 72, 72, 75, 78, 80, 84, 84, 85, 88, 88, 90, 93, 95, 99, 100, 100

(a) Use your calculator to determine the mean, the median, and the quartiles for this data set. Then, construct a simple box-and-whiskers (box plot) for this data set.

(b) What is the interquartile range of this data set? In theory, what percent of the data set should lie between the first and third quartiles? Is that true for this data set?

(c) What is the population standard deviation for this data set to the nearest tenth? How do you interpret the standard deviation?

(d) What percent of the scores were within one standard deviation of the mean? Within two standard deviations of the mean? Round your percents to the nearest percent and show your work.

Within One Standard Deviation of the Mean

Within Two Standard Deviations of the Mean

Sometimes data is grouped in a frequency chart. We still should be able to calculate the basic population parameters when the information is given in this form.

***Exercise* #3:** A small company has salaries for their 50 employees as given in the table below

(a) Find the mean and standard deviations of the salary range.

(b) What is the median of this data set? Why is the median considerably lower than the mean in this data set?

Salary (x_i)	Frequency (f_i)
25,000	5
32,000	21
45,000	14
58,000	7
75,000	2
120,000	1

(c) Does more or less than 50% of the data set fall within one standard deviation of the mean? Show the analysis that leads to your answer.

Although we have often concentrated on experimental studies where data is collected and means are found, many times we use statistics to represent results of a survey where we are interested in what **proportion** of a **population** share a certain characteristic. These proportions are most expressed as decimals, but sometimes are represented by fractions or percents.

***Exercise* #4:** A questionnaire went home to all juniors concerning their ability to bring and use mobile devices at school. The questionnaires constituted a **census** since all of the juniors were surveyed. Of the 742 juniors, 564 of them reported having web-enabled mobile devices. What was the population proportion for web-enabled devices? Express your answer as a decimal and as a percent.

***Exercise* #5:** The proportion of eggs that get cracked in a local egg handling facility is 0.023. If 2,500 dozen eggs are packaged in the factor per day, what should we expect to be the number of eggs cracked per day?

(1) 350 (3) 230

(2) 450 (4) 690

Name: _____ Date: _____

A2CC: Intro to Statistics

HOMEWORK

APPLICATIONS

1. Scientists randomly select ten groups from a population of men over 50 years old. They calculate the mean weights of each of these groups. The variability between these means can be best attributed to

   (1) measurement variability (3) induced variability

   (2) natural variability (4) sampling variability

   _____

2. Max and Daniel are measuring the amount of time it takes for a ball to roll down a ramp at different heights. For each trial, both Max and Daniel take turns rolling the ball and working the stop watch. They do this in order to quantify which of the following sources of variability?

   (1) measurement variability (3) induced variability

   (2) natural variability (4) sampling variability

   _____

3. Which of the following scenarios would be an attempt to quantify induced variability?

   (1) a phone survey of political preferences during election season
   (2) multiple random samples of products from an assembly line to check for defects
   (3) random assignment of people to a control group and a group taking a drug to lower cholesterol
   (4) recording the variability in the measurement of a soil sample's weight by the same machine

   _____

4. Which of the following research questions would involve collecting data through a survey?

   (1) Watching people exit a grocery store to see the percent who use reusable bags.
   (2) Assigning people to two groups to see the effect of a particular amount of sleep.
   (3) Calling people on the telephone to see if they will be voting in the upcoming election.
   (4) Dropping salt cubes into two different liquids to determine which dissolves faster.

   _____

5. In which of the following cases would an observational study be necessary as compared to an experimental study?

   (1) The study of how increased nutrient levels affect plant growth.
   (2) The study of how educational levels affect median household income.
   (3) The study of how a vaccine affects the percent of mice that get a particular disease.
   (4) The study of how noise level affects the sleep patterns of volunteers in a sleep study.

   _____

6. Which of the following formulas, written in summation notation, would represent the mean of the data set $\{x_1, x_2, ..., x_n\}$? Explain your choice.

(1) $\sum_{i=1}^{n} x_i$ (3) $n\sum_{i=1}^{n} x_i$

(2) $\frac{1}{n}\sum_{i=1}^{n} x_i^2$ (4) $\frac{1}{n}\sum_{i=1}^{n} x_i$ _____

7. The standard deviation of a population characteristics measures

(1) The difference between the maximum and minimum values.
(2) The difference between the third quartile and first quartile values.
(3) The average distance a data value is away from the mean.
(4) The average distance a data value is away from the median. _____

8. The interquartile range of the data set $\{4, 7, 10, 13, 18, 22, 30\}$ is

(1) 15 (3) 7

(2) 18 (4) 10

9. If 348 freshmen out of 622 have cell phones, then the population proportion, *p*, for freshmen cell phone ownership is

(1) 0.56 (3) 0.72

(2) 0.35 (4) 0.44 _____

10. If a population has 824 subjects, then about how many would have characteristics in the upper quartile?

(1) 412 (3) 368

(2) 280 (4) 206

11. In an experimental study, a lab wanted to divide volunteers into two groups to determine the effect of a particular phone app to help make people more punctual (on time). The 50 volunteers in the study will be assigned to either a group of 25 who use the app for a week or a group of 25 who do not use the app. The participants were asked to come to a lab to receive the app (or not) at 10:00 am on a Monday. Answer the following questions:

    (a) Why would those performing the study *not* want to assign the participants in the two treatments (groups) based on who showed up to the study session first?

    (b) Propose a way to use a random number table to generate a simple random selection that eliminates the bias that you discussed in part (a).

12. A school is tracking its freshmen attendance for the first marking period. Shown below is a table summarizing their findings for the 284 members of the freshmen class.

    (a) Find the mean and median number of days absent. Round your mean to the nearest tenth.

    (b) What is the population standard deviation for this data set? Round to the nearest tenth.

    (c) What proportion of the population that has an absenteeism greater than 4 days?

Days Absent (x_i)	Number of Students (f_i)
0	158
1	64
2	18
3	22
4	4
5	7
6	8
9	2
13	1

13. The heights of the 15 players on the Arlington boys' varsity basketball team are given below in inches.

66, 67, 68, 68, 70, 72, 72, 73, 74, 75, 75, 75, 76, 77, 79

(a) Find the mean and standard deviation of this data set. Use the population standard deviation. Round both to the nearest *tenth*.

(b) Determine the proportion of the population that falls within one standard deviation and within two standard deviations of the mean. State your values in decimal form.

One standard deviation from the mean:

Two standard deviations from the mean:

(c) Use the random number table for this lesson to pick a random sample of five players from this list. Do this by picking a random two digit column along the page. Scan down the column until you have picked 5 random integers that fall from 1 to 15. Write down your sample and calculate its mean.

14. If you were trying to conduct a survey of political preferences for likely voters in an upcoming election and decided to dial 1,000 randomly generated land-line phone numbers (not cell), why might this still introduce **bias** into the sampling?

15. Two groups of subjects were divided in an experimental study. One group was given a drug to help speed up their metabolism and result in weight loss. The other group was given a **placebo** (a pill that looks identical to the one given to the other group, but without the weight loss drug). After a month of the experiment, the weight loss of each individual in each of the two groups were measured. In general, people in the group who took the metabolism drug did lose more weight, although there were differences in the amount each lost. There are two main types of variability occurring in this study. Describe each type below in the context of this study.

Induced Variability

Natural Variability